IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.19, No.3, July 2025, pp. 235~246
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.98531

Enhancing Image Classification Performance Using Multi CNN Feature Fusion Method

Hizbullah Hamda*[1], Moh. Edi Wibowo[2]
[1]Master Program of Computer Science, FMIPA UGM, Yogyakarta, Indonesia
[2]Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia
e-mail: ***[1]hizbullah.hamda@mail.ugm.ac.id**, [2]mediw@ugm.ac.id

Abstract

Penelitian ini bertujuan untuk mengatasi tantangan umum dalam bidang pengenalan pola citra menggunakan Convolutional Neural Network (CNN), yang masih dihadapkan pada kompleksitas dan keterbatasan data citra. Mencapai akurasi tinggi sangat penting karena secara signifikan mempengaruhi efektivitas dan keberhasilan berbagai aspek. Meskipun teknologi deep learning, terutama CNN, menawarkan potensi untuk meningkatkan akurasi, namun masih terbatas pada kisaran 70 hingga 80% dalam mencapai tingkat akurasi yang diharapkan.

Dalam penelitian ini, dikembangkan metode fusion yang menggabungkan pre-trained model menggunakan teknik concatination untuk meningkatkan akurasi. Dengan memanfaatkan pre-trained model seperti ResNet50, VGG16, dan MobileNet-V2, yang kemudian disesuaikan dengan berbagai dataset dan teknik Cross-Validation berhasil mencapai peningkatan yang signifikan dalam akurasi.

Hasil penelitian ini menunjukkan peningkatan dalam akurasi model Fusion Multi CNN untuk berbagai dataset. Pada dataset Fashion MNIST berhasil mencapai akurasi 0,87840, sementara pada CIFAR-10 dan Oxford-102 dengan akurasi masing-masing sebesar 0,81260 dan 0,84004.

Kata kunci*— convolutional neural network (CNN), fusion fitur multiple CNN, cross validation, transfer learning, pre-trained model*

Abstract

This research aims to overcome general challenges in the field of image pattern recognition using a convolutional neural network (CNN), which is still faced with the complexity and limitations of image data. Achieving high accuracy is essential because it significantly influences the effectiveness and success of numerous areas. Although deep learning technology, especially CNNs, offers the potential to improve accuracy, it is still limited to the 70–80% range for achieving the expected level of accuracy.

In this research, a fusion method was developed that combines pre-trained models using concatenation techniques to increase accuracy. By utilizing pre-trained models such as ResNet50, VGG16, and MobileNet-v2, which were then adapted to various datasets and cross-validation techniques, researchers managed to achieve significant improvements in accuracy.

The results of this study show an improvement in the accuracy of the Fusion Multi-CNN model for various datasets. On the fashion dataset, MNIST managed to achieve an accuracy of 0.87840, while on CIFAR-10 and Oxford-102, the accuracy was 0.81260 and 0.84004, respectively.

Keywords— *convolutional neural network (CNN), fusion fitur multiple CNN, cross validation, transfer learning, pre-trained model*

Received June 1st,2025; Revised July 25th, 2025; Accepted July 31th, 2025

1. INTRODUCTION

The use of Convolutional Neural Networks (CNN) has brought significant advancements in image processing and pattern recognition [1], [2]. CNN achieved high accuracy in medical classification tasks, notably in osteoporosis detection using dental radiograph images [3]. CNNs have become one of the most successful approaches in image classification due to their ability to automatically extract hierarchical features [4]. However, CNNs still face challenges in dealing with the complexity and diversity of image classification data [5], [6].

An effective approach to improve CNN performance is by combining the feature extractions from multiple CNN models in a feature fusion technique [7]. Feature fusion leverages the strengths and unique characteristics of each model to enhance the accuracy and robustness of pattern recognition [8]. This technique has been proven effective in both image classification and healthcare applications, such as early detection of chronic heart failure, by retaining effective discriminative structures and reducing variance, thus producing more general and robust predictions [9], [10].

Fusion method is particularly useful in medical data analysis due to its ability to integrate various predictive models without significantly altering the original data, unlike methods such as data augmentation, which risk creating unrealistic samples [11]. Moreover, techniques such as dropout layers and feature selection must be applied carefully to medical data to avoid losing important information [12], [13].

Cross-validation techniques have become an effective method for evaluating and validating model performance, helping to avoid bias and overfitting [14], [15]. Through cross-validation, more reliable estimates of a model's generalization capability can be obtained [16].

This research aims to combine multiple CNN fusion techniques with cross-validation in image pattern recognition. This method is expected to enhance the performance and reliability of image classification systems through the use of multiple CNN ensembles optimized with cross-validation techniques.

2. METHODS

The main process in this research involves the preparation of the data used, the creation of pre-trained models, namely VGG16, ResNet-50, and MobileNet-v2, followed by feature extraction using these pre-trained models. The model fusion combines the extracted features from the three models. The resulting extracted features are then fed into softmax and random forest classifiers, along with the evaluation metrics used to assess the models.

2.1 Dataset

In this research, the image data used are selected from credible sources to ensure relevance to the research objectives. The chosen datasets encompass various categories of objects to ensure a comprehensive evaluation of the performance of the ensemble multiple CNN using the cross-validation method.

The main dataset used is Fashion MNIST, consisting of 28×28 pixel images representing various types of clothing and fashion accessories. This dataset contains 70,000 images, divided into 60,000 for training and 10,000 for testing. Fashion MNIST offers a sufficient variety of categories to test the model's ability to recognize and distinguish distinctive features, and provides a more complex challenge compared to the simple MNIST dataset.

Additionally, this research also utilizes the CIFAR-10 dataset, which consists of 60,000 color images of 32×32 pixels across 10 common object classes. This dataset is divided into 50,000 for training and 10,000 for testing. CIFAR-10 offers considerable diversity in its classes and is a standard in image recognition research, allowing for comparisons with previous studies.

The third dataset used is Oxford Flower 102, which has fewer samples compared to Fashion MNIST and CIFAR-10. This dataset is designed for the task of flower classification with

102 categories and significant variation in scale, pose, and lighting in the images. Oxford Flower 102 helps test the adaptability and effectiveness of the fusion model under diverse data conditions.

The goal of using these three datasets is to evaluate the reliability of the machine learning system in image recognition and to identify the extent to which the system can generalize its knowledge across different datasets. This is crucial for understanding the system's performance in various real-world situations and for testing the overall reliability and robustness of the models.

2.2 Preprocessing

Before image data can be used in the training and evaluation phases of the ensemble, preprocessing steps are applied to enhance data quality and reduce any potential noise in the images. This preprocessing includes normalizing pixel values, adjusting image sizes, and cleaning the data to remove unwanted noise. Adjusting the image size is a crucial step, especially when using pre-trained models, as each model typically requires specific input image sizes to function properly. Therefore, images that do not match the required size must be resized to be compatible with the model. This involves techniques such as cropping, padding, or scaling to ensure each image has a consistent size suitable for the model architecture. For instance, models like VGG16 or ResNet50 require images to be at least 32×32 pixels. Without these adjustments, images cannot be processed correctly by the model, potentially leading to performance degradation or even processing failures.

After preprocessing, the dataset is divided into two key subsets: training and testing. This division is carried out using cross-validation, where the dataset is split into multiple parts that are alternately used for training and testing the model. This approach ensures that the model is tested on diverse data that it has not seen during training, providing a more accurate measure of the ensemble's performance. These steps ensure that the image data used in ensemble learning is optimally processed, sufficiently varied, and ready to support the development of an ensemble model to address the research challenges.

2.3 Cross Validation

The cross-validation step will be applied to validate and optimize the performance of the ensemble multiple CNN on previously unseen data. This research uses the 5-Fold Cross-Validation technique to split the dataset into training and testing subsets. The use of 5-Fold Cross-Validation is illustrated in Figure 1.

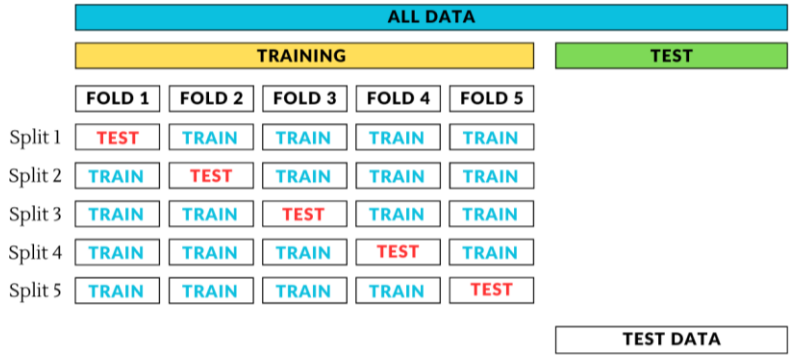

Figure 1 Cross validation process

In Figure 1, the use of the 5-Fold Cross-Validation technique to divide the dataset into training and testing subsets is shown. In 5-Fold Cross-Validation, the dataset is divided into five nearly equal subsets. Each subset takes turns being the testing dataset, while the other four subsets become the training dataset. This process is repeated five times, with each subset serving as the testing dataset once.

The Fashion MNIST dataset used in this research consists of 60,000 training data and 10,000 testing data. To implement the 5-Fold Cross-Validation method, the first step is to split

the training data into five smaller subsets. This splitting process results in subsets (folds) each containing 12,000 data points.

The main advantage of the cross-validation method is that it provides a more objective view of the model's generalization ability on unseen data during training. By ensuring that each sample is part of the testing dataset at least once, this method can produce more reliable estimates of model performance and address issues such as overfitting (the model being too well-suited to the training data) or underfitting (the model being too simple to capture patterns in the data).

2.4 Pre-trained Model

In this research, three pretrained models renowned for their effectiveness in image processing are utilized: VGG16, ResNet-50, and MobileNet-v2. Each model brings distinct capabilities to the ensemble, tailored to handle various aspects of complex image recognition tasks.

VGG16 features a relatively simple structure with 16 layers, characterized by successive convolutional layers using small (3×3) filters followed by max pooling. This architecture excels in extracting basic features such as edges, shapes, and textures, making it ideal as an initial feature extractor in the ensemble.

ResNet-50, on the other hand, embodies the residual network architecture designed to address performance degradation in deep networks. By introducing residual blocks, where layers not only learn new features but also differences from previous features, ResNet-50 can effectively handle deeper networks without suffering from diminishing gradients. This capability is crucial in combating issues like overfitting.

MobileNet-v2, developed for mobile and embedded applications, emphasizes efficiency without sacrificing accuracy. It utilizes depthwise separable convolutions to reduce computational costs while maintaining robust feature representation. Features like inverted residuals and linear bottlenecks further enhance performance in resource-constrained environments, such as mobile devices.

The selection of these pretrained models is based on their individual strengths and suitability for the task at hand. VGG16's reputation in image recognition, ResNet-50's resilience to gradient vanishing, and MobileNet-v2's efficiency and feature representation quality collectively contribute to a well-rounded ensemble approach.

Each pretrained model is rigorously tested across three distinct datasets to evaluate its performance comprehensively. For instance, adapting grayscale images from Fashion MNIST to RGB format ensures compatibility with the pretrained models while preserving the original information. This conversion process involves replicating grayscale information into three RGB channels, maintaining data integrity while aligning it for processing by the models.

By leveraging the strengths of VGG16, ResNet-50, and MobileNet-v2 in fusion methods, this research aims to enhance the reliability and performance of image recognition systems, particularly in handling complex image classification tasks across diverse datasets.

2.5 Fusion

The fusion method adopted in the ensemble multiple Convolutional Neural Network (CNN) is concatenation. Concatenation, also known as concatenation, is a robust approach to combining outputs from various models in an ensemble, allowing for richer and more complex information processing. The process of concatenation is illustrated in Figure 2.

In Figure 2, the concatenation process is depicted. In the ensemble multiple CNN setup, the results of concatenation are used to combine outputs from each CNN model in the ensemble, including VGG-16, ResNet-50, and MobileNet V2. After each model performs image classification in its final stage, the outputs from these final layers are taken as feature representation vectors of the processed images. These vectors are then concatenated into a single long vector that encompasses information from all models in the ensemble.

At this stage, the concatenated information is processed to generate a final prediction. This unified output integrates all the information learned by the three models, aiming to improve prediction accuracy compared to using each model separately.

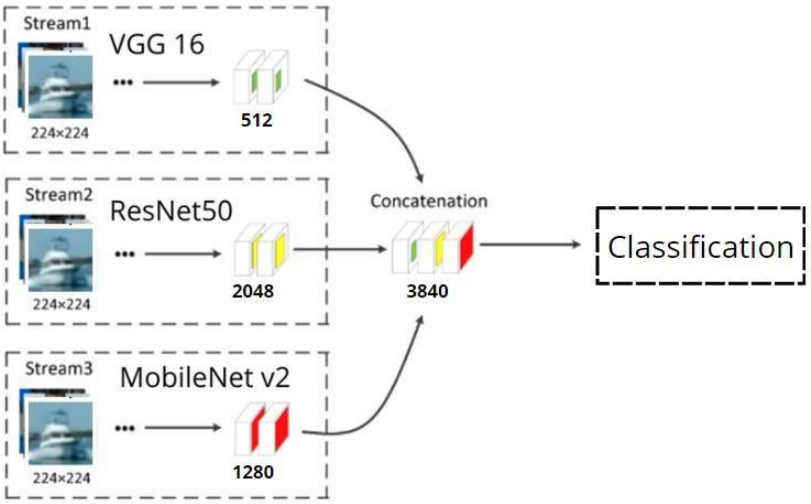

Figure 2 Concatenation

2.6 Classification

In this research, two methods are employed for classification: ensemble multiple Convolutional Neural Networks (CNN) and Random Forest. In the classification phase using ensemble multiple CNNs that have been designed and pretrained, the process involves the outputs from each model in the ensemble. Each model generates its own classification predictions based on the extracted features. These predictions are then combined through concatenation of output vectors into a single long vector. Subsequently, this concatenated vector undergoes processing through a final classification layer comprising fully connected layers and a softmax layer to produce the final prediction.

The results from the classification layer yield a probability distribution for each class, and decision-making is performed by applying a threshold to the probability distribution. The class with the highest probability is chosen as the final classification for the given image. Using this approach, ensemble multiple CNNs integrate information from various models to generate a robust final prediction. Careful decision-making and evaluation provide guidance in understanding the ensemble's performance in image classification tasks, facilitating the achievement of research objectives by integrating predictions from multiple CNNs to produce reliable and consistent classification outcomes.

The second method employed for classification is using Random Forest. The combined feature vector from multiple CNNs serves as input for the Random Forest classification model. The Random Forest model discerns complex patterns within this feature vector and classifies them according to predefined class labels. The choice of Random Forest as a comparative algorithm after the fusion process of several pretrained CNN models is based on several important considerations.

Firstly, we acknowledge Random Forest's ability to address overfitting issues, which are often a primary concern when using complex models on relatively small datasets like CIFAR-10. With a balanced trade-off between variance and bias, Random Forest helps minimize the risk of overfitting that may occur after merging different CNN models. Additionally, the Random Forest algorithm is well-known for its capability in handling multiclass data, which aligns with the nature of the CIFAR-10 dataset consisting of 10 different classes.

2.7 Performance Evaluation

The performance evaluation method adopted in this research aims primarily to measure how well the ensemble multiple CNN results obtained using cross-validation approach in image classification systems. In this context, several classification evaluation metrics are used to provide a more comprehensive overview of the system's ability to recognize and classify objects in images, such as accuracy, precision, recall, and F1-score.

Accuracy is used to measure how well the system can classify data correctly overall. It is calculated as the ratio of the number of correctly classified data to the total number of data in the dataset. Furthermore, precision and recall are used to provide deeper insights. Precision measures the percentage of true positive results out of all positive predictions made by the system. It assesses the level of accuracy in identifying a specific class. On the other hand, recall measures how well the system can find or detect all positive instances in the dataset.

Lastly, the F1-score is a metric that combines precision and recall. It provides a balance between precision and recall, and is particularly useful when the class distribution is imbalanced or when the positive class appears in limited quantities.

2.8 Testing Scheme

In the conducted research, a testing scenario has been designed to evaluate four different machine learning models. The first model used is VGG-16, the second is ResNet-50, the third is MobileNetV2, and the fourth model is a fusion model combining the three previous models. The testing scheme for the models can be seen in Figure 3.

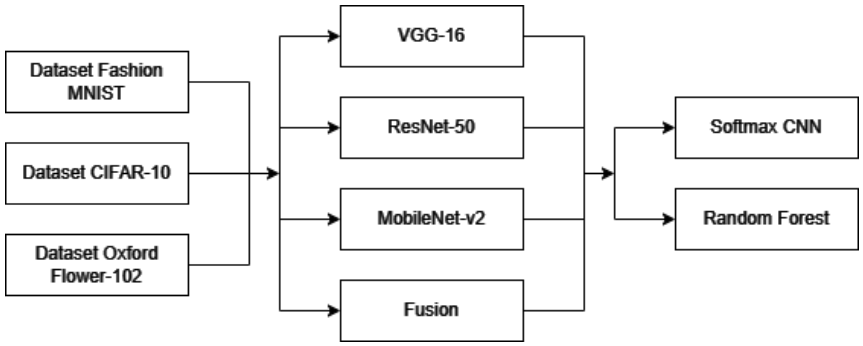

Figure 3 Test scheme

In Figure 3, the testing scheme illustrates the evaluation of the pre-trained models: VGG-16, ResNet-50, MobileNetV2, and the fourth model, which is a fusion of these three models. The fusion process involves testing different combinations of the three pre-trained models (VGG-16, ResNet-50, and MobileNetV2) in various orders. After training these models, they proceed to the classification stage. In classification, each model is paired with two types of classification mechanisms: softmax CNN and the Random Forest classification algorithm. By using these two classification methods, the research not only compares different machine learning models but also evaluates how different classification methods affect the performance of each model.

Following the classification process, the research generates standard classification matrices for each model (model evaluation). Using data from these classification matrices, comparisons are made between the results obtained from each model. This comparison involves not only comparing performance metrics but also conducting deeper analyses of why certain models perform better or worse under specific conditions. This process is crucial for understanding the strengths and limitations of each tested model.

The final stage of this testing scenario involves comprehensive evaluation and drawing conclusions based on all the data and analyses conducted. This evaluation includes a review of all findings gathered during the research, using them to derive meaningful conclusions about the

performance and practical applications of the tested machine learning models. These conclusions aid not only in understanding the current landscape of image classification algorithms but also in guiding future research and application development based on these findings.

3. RESULTS AND DISCUSSION

This section will include the results of the model with different dataset and most optimal sequence for combining different CNN models.

3.1 Model with Dataset Fashion MNIST

In this research, training and testing of several models using the MNIST dataset have been conducted. All four models use two types of classification, namely Softmax CNN and Random Forest.

Table 1 Results of the model using the Fashion MNIST dataset.

Model	Softmax CNN Accuracy		Random Forest Accuracy		
	Val	Test	Val	Test	Training Time
VGG 16	0,81657	0,81210	0,81008	0,80950	28,65
ResNet-50	0,85468	0,84570	0,83205	0,82900	266,78
MobileNetV2	0,80372	0,81040	0,77552	0,77300	163,88
Fusion (Proposed Method)	0,88033	0,87840	0,83792	0,84140	342,43

Based on the test results in Table 1, the softmax CNN classification model achieved the highest accuracy with the fusion pretrained model, achieving a validation accuracy of 0,88033, and testing accuracy of 0.87840. Overall, the use of pretrained models indicates that fusion performs better in accuracy compared to the other three pretrained models. However, the training time required by the fusion model in both classification models is the longest, taking 342.43 seconds. In contrast, the pretrained VGG 16 model required the shortest training time, only 28.65 seconds, achieving a validation accuracy of 0.81657, and testing accuracy of 0.81210. Meanwhile, the Random Forest classification model achieved the highest accuracy in the fusion model, with validation and testing accuracies of 0,83792 and 0,84140, respectively.

These results indicate that while the fusion of pretrained models shows promising performance in terms of accuracy, it also comes with a trade-off of longer training times compared to individual pretrained models like VGG 16. Random Forest, despite its very high training accuracy, shows slightly lower accuracy in validation and testing compared to the model in the softmax CNN classification. This comparison highlights the importance of not only accuracy but also considerations such as computational efficiency and generalization across different stages of model development and testing.

3.2 Model with Dataset CIFAR-10

Next, experiments are conducted using the CIFAR-10 dataset, which is more complex than the Fashion MNIST dataset. CIFAR-10 consists of color images categorized into 10 different object classes, whereas Fashion MNIST comprises grayscale images categorized into 10 different clothing categories.

The softmax CNN classification model achieved the highest accuracy with the fusion pretrained model, achieving a validation accuracy of 0,81650, and testing accuracy of 0,81260. Similarly, the Random Forest model also achieved the highest accuracy with the fusion pretrained model, with validation accuracy of 0,68960, and testing accuracy of 0,68760. Overall, the use of pretrained models indicates that fusion performs better in accuracy compared to the other three pretrained models. However, the training time required by the fusion model in both classification

setups is the longest, taking 342.43 seconds. In contrast, the pretrained VGG 16 model required the fastest training time, only 58.98 seconds.

Table 2 Results of the model using the CIFAR-10 dataset.

Model	Softmax CNN Accuracy		Random Forest Accuracy		
	Val	Test	Val	Test	Training Time
VGG 16	0,71936	0,71620	0,66694	0,66290	28,65
ResNet-50	0,75558	0,76260	0,64618	0,65150	266,78
MobileNetV2	0,43798	0,43360	0,39002	0,39560	163,88
Fusion (Proposed Method)	0,81650	0,81260	0,68960	0,68760	342,43

These results in Table 2 highlight the trade-offs between accuracy and training time across different models and datasets. The fusion of pretrained models consistently shows strong performance in accuracy metrics, especially in complex datasets like CIFAR-10. However, this approach requires longer training times due to the integration and combination of multiple models' features. Conversely, models like VGG 16 demonstrate faster training times but slightly lower accuracy compared to the fusion approach. Random Forest, shows lower but still competitive validation and testing accuracies.

3.3 Model with Dataset Oxford-102

The final dataset used is the Oxford Flower 102 dataset, which includes 102 distinct classes of flowers. This dataset is widely used for fine-grained image classification tasks due to its diverse floral categories and high-resolution images, making it suitable for detailed visual recognition studies.

Table 3 Results of the model using the Oxford-102 Flower dataset.

Model	Softmax CNN Accuracy		Random Forest Accuracy		
	Val	Test	Val	Test	Training Time
VGG 16	0,63338	0,60928	0,45251	0,42125	28,65
ResNet-50	0,078290	0,07692	0,16092	0,16728	266,78
MobileNetV2	0,77476	0,75336	0,63786	0,59829	163,88
Fusion (Proposed Method)	0,85207	0,84004	0,64180	0,61904	342,43

The results shown in Table 3 of using the Oxford-102 dataset, softmax CNN classification model achieved the highest accuracy with the fusion pretrained model, achieving a validation accuracy of 0,85207, and testing accuracy of 0,84004, with a training time of 61.63 seconds. RF model achieved the highest accuracy with the fusion model, validation accuracy of 0,64180, and testing accuracy of 0,61904, with a training time of 58.58 seconds. Overall, the use of pretrained models indicates that fusion performs very well in terms of accuracy. However, the training time required by the fusion model in both classification models is the longest, at 61.63 seconds, whereas the pretrained ResNet50 model required the shortest training time, only 16.10 seconds, albeit with the lowest accuracy compared to other models.

ResNet50 demonstrates the shortest training time among the tested pretrained models, clocking in at just 16.10 seconds. This efficiency in training duration can be attributed to ResNet50's architecture, which includes residual connections that aid in mitigating the vanishing

gradient problem, thereby potentially speeding up convergence during training. However, despite its quicker training time, ResNet50 may exhibit lower accuracy compared to other models.

The lower accuracy could stem from several factors. Firstly, ResNet50, while deep and effective, might require more epochs or tuning of hyperparameters to achieve peak performance comparable to the other models. Secondly, its complexity and depth may necessitate more computational resources during training, even if it converges faster initially. Lastly, the model's design and architectural nuances, such as the way residual connections influence feature learning, could also impact its final accuracy output.

3.4 Concatenation

This research adopts a sequential concatenation approach to evaluate the most optimal sequence for combining different CNN models. A total of 6 (six) models are constructed with different concatenation orders to compare the performance of each sequence.

Table 4 Concatenation model with dataset MNIST

Model	Softmax CNN Accuracy		Random Forest Accuracy	
	Val	Test	Val	Test
VGG,ResNet50, MobileNetV2	0,88033	0,87840	0,83792	0,84140
VGG, MobileNetV2, ResNet50	0,87333	0,86790	0,83758	0,83840
ResNet50,VGG, MobileNetV2	0,87600	0,87870	0,83950	0,83970
ResNet50, MobileNetV2,VGG	0,87350	0,87090	0,84008	0,84010
MobileNetV2,VGG, ResNet50	0,87358	0,87130	0,84142	0,83900
MobileNetV2, ResNet50,VGG	0,87633	0,87420	0,83933	0,83650

In Table 4, the results of concatenating models using the MNIST dataset are presented. From the conducted tests, the combination that achieved the highest validation accuracy in the CNN classification model was the sequence VGG, ResNet-50, MobileNet15 with an accuracy of 0.88033 and the hishest test accuracy is 0.87870 with combination of ResNet50,VGG and MobileNetV2. For the RF classification model during validation, the combination of MobileNetV2, VGG, and ResNet50 achieved the highest accuracy of 0.84142. In the testing phase, the combination of models achieved the highest accuracy of 0.84140. Overall, the combination of models that consistently performed lower across each model was the combination of VGG, MobileNetV2, and ResNet5.

Table 5 Concatenation model with dataset CIFAR-10

Model	Softmax CNN Accuracy		Random Forest Accuracy	
	Val	Test	Val	Test
VGG,ResNet50, MobileNetV2	0,81650	0,81260	0,68960	0,68760
VGG, MobileNetV2, ResNet50	0,80970	0,79780	0,68550	0,68360
ResNet50,VGG, MobileNetV2	0,81100	0,79950	0,68630	0,67820
ResNet50, MobileNetV2,VGG	0,81350	0,80140	0,69160	0,67990
MobileNetV2,VGG, ResNet50	0,82050	0,80700	0,68690	0,68100
MobileNetV2, ResNet50,VGG	0,81280	0,80350	0,68550	0,67990

In Table 5, the results of concatenating models using the CIFAR-10 dataset are shown. From the conducted tests, the combination of MobileNetV2, VGG, and ResNet50 achieved the highest validation accuracy in the CNN classification model, with a value of 0.82050. For the highest testing accuracy, the combination of VGG, ResNet50, and MobileNetV2 achieved a score of 0.81260.

Concatenation testing on the RF in the validation phase, the combination of ResNet50, MobileNetV2, and VGG achieved the highest accuracy of 0,69160. Meanwhile, in the testing phase, the combination of VGG,ResNet50, MobileNetV2 achieved the highest accuracy of 0,68760. Overall, the combination yielding the lowest results across almost every model was consistently VGG, MobileNetV2, and ResNet50.

Table 6 Concatenation model with dataset Oxford-102

Model	Softmax CNN Accuracy		Random Forest Accuracy	
	Val	Test	Val	Test
VGG,ResNet50, MobileNetV2	0,85207	0,84004	0,64180	0,61904
VGG, MobileNetV2, ResNet50	0,85696	0,83516	0,63691	0,61904
ResNet50,VGG, MobileNetV2	0,85941	0,85714	0,62958	0,62393
ResNet50, MobileNetV2,VGG	0,86063	0,85103	0,64547	0,61416
MobileNetV2,VGG, ResNet50	0,85207	0,84981	0,65036	0,60805
MobileNetV2, ResNet50,VGG	0,85696	0,84737	0,65525	0,60927

In Table 6, the results of concatenating models using the Oxford-102 dataset are presented. From the softmax, the combination with the order ResNet50, MobileNetV2, and VGG achieved the highest validation accuracy of 0.86063. Meanwhile, the highest testing accuracy was attained by the combination of ResNet50, VGG, and MobileNetV2 with a score of 0.85714.

For the RF classification model, in the validation phase, the combination of MobileNetV2, ResNet50, and VGG achieved the highest accuracy of 0.65525. Similarly, in the testing phase, the combination of ResNet50, VGG, and MobileNetV2 achieved the highest accuracy of 0.62393.

Based on the results of the concatenation model testing in Tables 4, 5, and 6, both the classifiers used in this study are designed to work independently with the provided input features. They do not consider the order of concatenation during the classification process as they only use the final features after extraction. These methods are permutation-invariant, meaning that as long as the features provided are the same, the order of concatenation does not affect the final output. Since the classifiers in this study work with feature vectors generated after extraction, they do not concern themselves with how these features are generated or ordered, as long as the features are relevant and contain enough information to distinguish between different classes. Although there are numerical differences observed, these are not a result of the order of feature concatenation but rather the natural variation occurring in experiments due to stochastic variations in the data.

4. CONCLUSIONS

In conclusion, the CNN Fusion model demonstrates superior performance compared to RF across all three datasets, achieving significant accuracies: 0.87840 on Fashion MNIST, 0.81260 on CIFAR-10, and 0.84004 on Oxford-102, consistently outperforming other pretrained models. The results from the concatenation order experiments indicate that the sequence of models that the classifiers used in this study are independent of the order of feature concatenation. These permutation-invariant methods focus solely on the final feature vectors after extraction,

ensuring that the sequence of feature combination does not affect the classification output. Although numerical differences are observed, they stem from natural experimental variations due to stochastic factors in the data, not from the feature concatenation order.

While the research has shown promising results, future developments could explore ensemble learning techniques or model-level fusion to further enhance accuracy and prediction reliability by harnessing the strengths of multiple models across datasets. Additionally, alongside standard evaluation metrics, incorporating interpretability evaluations such as model interpretability analysis or assessments of prediction reliability in challenging scenarios could provide deeper insights into model robustness and real-world applicability.

REFERENCES

[1] J. Butdee, W. Kongprawechnon, H. Nakahara, N. Chayopitak, C. Kingkan, and R. Pupadubsin, "Pattern Recognition of Partial Discharge Faults Using Convolutional Neural Network (CNN)," in *2023 8th International Conference on Control and Robotics Engineering*, Institute of Electrical and Electronics Engineers (IEEE), Jun. 2023, pp. 61–66. doi: 10.1109/iccre57112.2023.10155616.

[2] A. Tayal, J. Gupta, A. Solanki, K. Bisht, A. Nayyar, and M. Masud, "Correction to: DL-CNN-based approach with image processing techniques for diagnosis of retinal diseases," in *Multimedia Systems*, Springer Science and Business Media Deutschland GmbH, 2021, pp. 1417–1438. doi: 10.1007/s00530-021-00791-9.

[3] K. Hidjah, A. Harjoko, M. Edi Wibowo, and R. Ratna Shantiningsih, "Periapical Radiograph Texture Features for Osteoporosis Detection using Deep Convolutional Neural Network," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, pp. 223–232, 2022, doi: 10.14569/IJACSA.2022.0130127.

[4] D. Kollias and S. Zafeiriou, "Exploiting Multi-CNN Features in CNN-RNN Based Dimensional Emotion Recognition on the OMG in-the-Wild Dataset," *IEEE Trans Affect Comput*, vol. 12, no. 3, pp. 595–606, Jul. 2021, doi: 10.1109/TAFFC.2020.3014171.

[5] B. Paul and S. Phadikar, "A hybrid feature-extracted deep CNN with reduced parameters substitutes an End-to-End CNN for the recognition of spoken Bengali digits," *Multimed Tools Appl*, 2023, doi: 10.1007/s11042-023-15598-1.

[6] Z. Yu, J. Tang, and Z. Wang, "GCPS: A CNN Performance Evaluation Criterion for Radar Signal Intrapulse Modulation Recognition," *IEEE Communications Letters*, vol. 25, no. 7, pp. 2290–2294, Jul. 2021, doi: 10.1109/LCOMM.2021.3070151.

[7] Wu Zuobin, Mao Kezhi, and Gee -Wah Ng, "Effective feature fusion for pattern classification based on intra-class and extra-class discriminative correlation analysis," in *20th International Conference on Information Fusion (Fusion)*, 2017, pp. 1–8. doi: 10.23919/ICIF.2017.8009795.

[8] R. Laroca, L. A. Zanlorensi, V. Estevam, R. Minetto, and D. Menotti, "Leveraging Model Fusion for Improved License Plate Recognition," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, Springer Science and Business Media Deutschland GmbH, 2024, pp. 60–75. doi: 10.1007/978-3-031-49249-5_5.

[9] N. Gawande, D. Goyal, and K. Sankhla, "Improved Deep Learning and Feature Fusion Techniques for Chronic Heart Failure," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 17s, pp. 67–80, 2024, [Online]. Available: www.ijisae.org

[10] P. Deepan, "Fusion of Deep Learning Models for Improving Classification Accuracy of Remote Sensing Images," *JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES*, vol. 14, no. 5, Oct. 2019, doi: 10.26782/jmcms.2019.10.00015.

[11] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," Dec. 2017, [Online]. Available: http://arxiv.org/abs/1712.04621

[12] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," 2014.

[13] I. Guyon and A. M. De, "An Introduction to Variable and Feature Selection André Elisseeff," 2003.

[14] R. Malhotra and S. Meena, "Empirical Validation of cross-version and 10-fold cross-validation for Defect Prediction," in *Proceedings of the 2nd International Conference on Electronics and*

Sustainable Communication Systems, ICESC 2021, Institute of Electrical and Electronics Engineers Inc., Aug. 2021, pp. 431–438. doi: 10.1109/ICESC51422.2021.9533030.

[15] M. Yüzkat, H. O. Ilhan, and N. Aydin, "Multi-Model CNN Fusion for Sperm Morphology Analysis," *Comput Biol Med*, vol. 137, pp. 1–12, Oct. 2021, doi: 10.1016/j.compbiomed.2021.104790.

[16] Z. A. Sejuti and M. S. Islam, "A hybrid CNN–KNN approach for identification of COVID-19 with 5-fold cross validation," *Sensors International*, vol. 4, pp. 1–11, Jan. 2023, doi: 10.1016/j.sintl.2023.100229.